**Using “Big Data” to forecast migration - A tale of high expectations, promising results and a long road ahead**

Jasper Tjaden, Andres Arau, Muertizha Nuermaimaiti, Imge Cetin, Eduardo Acostamadiedo, Marzia Rango

**Act 1 - High expectations**

Data is the new oil, they say. ‘*Big* Data’ is even bigger than that. The “data revolution” will help solve societies’ problems and help governments adopt better policies and run more effective programs. In [the migration field](https://migrationdataportal.org/themen/big-data-migration-and-human-mobility), Big Data (or more accurately: digital trace data) is seen as a powerful tool to improve migration management processes ([visa applications](https://www.gmfus.org/publications/ai-digital-identities-biometrics-blockchain-primer-use-technology-migration-management); [asylum decision](https://www.bamf.de/SharedDocs/Anlagen/DE/Digitalisierung/broschuere-digitalisierungsagenda-2020.pdf?__blob=publicationFile&v=9) and [geographic allocation of asylum seeker, facilitating integration](https://immigrationlab.org/2018/05/26/switzerland-launches-program-test-ai-refugee-integration/), “[smart borders](https://ec.europa.eu/home-affairs/what-we-do/policies/borders-and-visas/smart-borders_en)” etc.)[[1]](#footnote-1)

Forecasting migration is one particular area where big data seems to excite data nerds (like us!) and policymakers alike. If there is one way that big data has already made a difference, it is its ability to bring different actors together, data scientists out of their basements and policy makers to sit through countless slides with numbers, tables and graphs. [Traditional migration data sources](https://migrationdataportal.org/themes/migration-data-sources), like censuses, administrative data and surveys, have never quite managed to incite the same level of excitement and policy momentum.

Many EU countries are currently heavily investing in new ways to forecast migration. Large numbers of asylum seekers in 2014, 2015 and 2016 strained the capacity of many EU governments. Better forecasting tools are meant to help governments prepare in advance.

A [recent European Migration Network study](https://migrationnetwork.un.org/events/11-2020/deep-dive-migration-40-forecasting) showed that 10 out of 22 EU governments that responded to the ad-hoc query said they make use of forecasting methods, many using open source data for “early warning and risk analysis” purposes. The 2020 European Migration Network conference was dedicated entirely to the theme of forecasting migration with hosting [more than 15 expert presentation](https://www.emn.at/en/national-emn-conference-2020-forecasting-the-future-of-global-migration/)s. The recently proposed EU Migration Pact outlines a “[Migration Preparedness and Crisis Blueprint](https://ec.europa.eu/info/sites/info/files/commission_recommendation_on_an_eu_mechanism_for_preparedness_and_management_of_crises_related_to_migration_migration_preparedness_and_crisis_blueprint_0.pdf)” which “should provide timely and adequate information in order to establish the updated migration situational awareness and provide for early warning/forecasting, as well as increase resilience to efficiently deal with any type of migration crisis.” (p. 4) The European Commission is currently finalizing a [feasibility study](https://migrationnetwork.un.org/events/11-2020/deep-dive-migration-40-forecasting) on the use of artificial intelligence for predicting migration to the EU; Frontex – the EU Border Agency – is scaling up efforts to [forecast irregular border crossings](https://migrationnetwork.un.org/events/11-2020/deep-dive-migration-40-forecasting); EASO – the EU Asylum Support Office – is devising a composite “[push-factor index](https://publications.iom.int/books/migration-west-and-north-africa-and-across-mediterranean-chapter-8)” and experimenting with [forecasting asylum-related migration flows using machine learning](https://arxiv.org/abs/2011.04348) and data at scale. In Fall 2020 – during Germany’s EU Council Presidency – the German Interior Ministry organized a workshop series around [Migration 4.0](https://migrationnetwork.un.org/migration-digitalization) highlighting the benefits of various ways to “digitalize” migration management. At the same time, the EU is investing substantial resources in migration forecasting research under its Horizon2020 programme including [QuantMig](https://www.southampton.ac.uk/quantmig/index.page), [ITFLOWS](https://cordis.europa.eu/project/id/882986/de), and [HumMingBird](https://hummingbird-h2020.eu/).

Is all this excitement warranted?

Yes, it is.

**Act 2 - Promising results**

Indeed, there are a growing number of studies out there that hint at the potential of big data forecasting on migration. Let’s take two prominent examples: the use of Google Trends and Facebook networks.

**Google Trends**

Google Trends is a powerful and publicly accessible tool for online search data. Google Trends is a good proxy for what over a billion users worldwide are curios about (i.e. searching for on the Google search engine).[[2]](#footnote-2) Migrants may use the internet to prepare for a journey (and also during the journey). This makes search data may be a potentially effective way of gaining insight on movement of migrants. There have been several studies exploring the use of Google Trends for predicting migration flows[[3]](#footnote-3).

Google Trends data has been used for [forecasting U.S migration](https://www.microsoft.com/en-us/research/uploads/prod/2019/03/ForecastMigration_www19.pdf), [forecasting migration to the OECD](http://conference.iza.org/conference_files/WoLabConf_2018/22651.pdf)  as well as [forecasting Latin America migration to Spain](https://core.ac.uk/download/pdf/294832951.pdf), among others.

The [Pew Research Center](https://www.pewresearch.org/) used searches of arrival ‘*destination*’ countries such as ‘Greece’ and ‘Italy’ among Arabic speakers in Turkey to track asylum seekers from the Middle -East (mainly from Syria and Iraq) to EU countries. Google searches were then compared to monthly asylum applications in Greece and found that Google searches for destination country for migrants were correlated with the actual migration data. Andre Groger, Tobias Heidland, and Marcus Bohme used Google Trends search keywords that were semantically linked with the word “migration” as a means to improve the performance of migration models. They conclude that Google Trends is a ‘novel’ way for measuring the intent to migrate and a better way to get real-time predictions of migration flow. Research looking at [Google Trends queries and migration flows from Latin America to Spain](https://core.ac.uk/download/pdf/294832951.pdf) find that online searches are correlated with records of past migration flows.

Researchers and data scientist everywhere are discovering the many potential ways they can utilize this publicly offered analytical tool. A number of studies find encouraging results, however, there there are still a lot of limitations before this tool can be used by policymakers.

Based on these promising case studies (see above), we (IOM’s GMDAC) set out to build a model that works globally and between all countries in the world. Digging into the data showed the many complications involved in using online search data such as how to technically query the data, the language settings, the specific terms and the data used for “ground truthing” (i.. comparing searches with official migration data). [[4]](#footnote-4)

During our 5-month research, we gradually and iteratively tried to find the best approach for each country pair in the world. Google searches appear to be similar to official migration statistics for some migration corridors (see Figure 1 below), but not others (see Figure 2 below).

We found that there was no ‘universal’ approach at the global scale and each bilateral migration seems to have its unique query approach.

Aspects like language settings, particular search term selection and other query related methodologies become a key point on how one can efficiently *optimize* the use of Google Trends.

Figure 1

Chart, line chart

Description automatically generated

The plot above shows the (pretty strong) correlation between the Google Trends Index and OECD migration inflows (our ground truthing data) between Syria and Sweden. We can see a relatively good relationship with the historic data.

But let’s not get too excited too soon. We also found many unique bilateral migration corridors which showed no clear relationship between google and migration data such as China to Korea (see Figure 2) and Iraq to Canada.

Figure 2

Chart, line chart

Description automatically generated

Granted, 5 month research won’t answer all our questions. Going forward we will need to take into account more context information (internet access, literacy rates, socio-economic development, geographical proximity, among others) to improve the model and understand why online searches are useful in some cases but not others. Up to now, the exercise showed the many pitfalls involved in using “big data” for migration forecasting and underlines the need for realistic expectations among policy makers.

**Facebook Connectedness Index**

With over [1.69 billion users around the globe](https://www.statista.com/statistics/490424/number-of-worldwide-facebook-users/#:~:text=This%20statistic%20shows%20the%20number,from%201.34%20million%20in%202014.) , Facebook has definitely become a very popular communication channel between individuals. Can it be used to forecast migration? Facebook data has been used to [monitor stocks of migrants](https://www.jstor.org/stable/26622775?seq=1) globally, to [estimate the increase of Venezuelan migrants](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224134) in Colombia and Spain or to assess the [cultural assimilation of Mexican immigrants in the U.S.](https://dl.acm.org/doi/abs/10.1145/3308558.3313409) In August 2020, Facebook made available data on friendship networks, the [Social Connected Index](https://dataforgood.fb.com/tools/social-connectedness-index/) measures the frequency and density of friendship and social ties around the world. For now, we only have a snapshot of networks at one point in time. If this data were available over time, changes in cross-country contacts could possibly be used to forecast global migration patterns. First preliminary results appear encouraging (see Figure 2 below).

We combined [Facebook’s SCI](https://dataforgood.fb.com/docs/social-connectedness-index-methodology/) with the [United Nations Population Divisions data on international migrant stock](https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp) explore whether this data has the potential to predict migration occurring from one country to other. The idea and approach are very simple: we used Facebook’s Connectedness Index between countries as a proxy for migration networks (social ties) and tested its correlation with the bilateral stock of migrants from the same country pair. The assumption is that changes in the networks across countries goes hand in hand with changes in international migration patterns.

Indeed, as Figure 2 shows, more mutual stock of migrants between countries correlates with a higher probability of Facebook friendship links between users in both locations. A combination of countries with a large migrant population like Mexico-United States or Morocco-Spain have a greater Connectedness Index than other combination of countries with fewer stock of bilateral migration like India-Argentina or Nigeria-Norway.

Figure 3

Chart, scatter chart

Description automatically generated

To further explore this relationship, we then tested if the positive correlation holds between migration stocks from different regions of the world and different income levels of countries. We find that 1 percent increase in networks is associated with a 0.7 percent increase in stocks. This is good news for forecasting. As more data will become available over time, a large increase in network size among countries may indicate an increase in migration between the countries before official traditional data sources will be available.[[5]](#footnote-5).

**Act 3 - A long road ahead**

So far, so promising. Yet, it will be a long way until forecasts based on this digital data will meet policy makers’ expectations.

The [Google Flu case](https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/) is a good example to explain why.

In 2009, [researchers from Google](https://www.nature.com/articles/nature07634) used google searches to “nowcast” the flu based on people’s searches. The paper – published in *Nature* - demonstrated that search data, if properly tuned to the flu tracking information from the Centers for Disease Control and Prevention, could produce accurate estimates of flu prevalence two weeks earlier than the U.S. Center for Disease Control’s data—turning the digital refuse of people’s searches into potentially life-saving insights. In 2014, a number of renowned researchers showed that after initial success the google flu tool “[failed spectacularly](https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf)” by missing the peak of the flu season by 140 percent. Instead of a “big data revolution” the authors claim “it is time we were focused on an “all data revolution”, where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.”

The Google Flu example is not a story of failure, but perhaps of moderation and complementarity. Google did not save the day, but it became a mainstream tool that the CDC now uses in combination with other data sources. Will we see the same for the case of forecasting migration? Google Flu was released 12 (!) years ago. The migration community has just now started to explore big data and all sorts of digital data sources for forecasting purposes.

Additional momentum from the policy side is helpful, but there is a risk of frustration when exceedingly high expectations are not met quickly. We should not get ahead of ourselves and accept that we are in an exploratory stage. It may take more than a decade for governments and national statistical offices to mainstream digital data as one instrument among others.

There are many more hurdles ahead. Compared to the Flu example, the migration community is facing several additional challenges: First, what people search when they have the flu is more specific than when they want to migrate. Migration is a complex social, political, and economic process rather than a clear physical symptom like the flu. Second, the Google Flu example is based in the US. There are many countries around the world with limited internet connectivity. Third, the US has good data on actual flu cases while the state of global migration data, however, is quite [poor](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5508757/). Fourth, regarding the digital trace data, in most of the cases, developed countries has exhaustive data than other developing countries. To test whether digital data can deliver accurate forecasts, it must first be compared first with reality in the form of accurate “traditional” data. This “ground truthing” is difficult as many countries in the world do not even systematically compile information on who is coming into their countries and who is leaving on an annual or monthly basis. Most countries in the world do not report that data to the UN even when it does exist.

Another general issue with Big Data is that it is user data from private companies. Facebook, for example, is less popular nowadays among younger cohorts. Changes in the user base may have dramatic implications for the ability to use this data for forecasting activities.

**Epilog: Patience and realism**

One likely trend that we will see is that results from big data forecasts will be combined or cross-validated by experts. As previous research shows there is no “one size fits all” method for migration forecasting, but each method has its own set of strengths and weaknesses. [Expert opinion on future migration flows](https://gmdac.iom.int/assessing-immigration-scenarios-european-union-2030) can be valuable as big data may be big, but it is also “thin” meaning that other context data is often lacking.

On the sidelines of a conference, Teddy Wilkin from EASO was sharing this story: “When my analysts detected a spike in searches for Italy among Nigerians, the machine took this at face value. In reality, we found that searches for Italy correlated with Champion League soccer games involving Italian teams”. According to Teddy, EASO does not let policy makers into the “machine room”. Results from forecasts are first contextualized and qualitatively analyzed by experts and then presented to policymakers alongside other sources.

It is true that data scientists and policymakers have never had so many interactions on this issue, yet they are still worlds apart. We have a long way to go before digital data become a reliable mainstream forecasting tool, but our experiences in just the last year have been very encouraging. The momentum is good, but it needs to sustain. Our hope is that policymakers and donor organizations don’t lose interest after the first gold rush. It is patience, realism and persistence that will allow the migration community to get the most value out of big data forecasting.

1. (Visit the [BigData4Migration Alliance](https://data4migration.org/) and the Global Migration Portal’s thematic page on “[big data, migration and human mobility](https://migrationdataportal.org/themen/big-data-migration-and-human-mobility)” for an overview of how big data has been used in the field of migration). [↑](#footnote-ref-1)
2. Google Trends is an index that Google has normalized and indexed to shows the relative popularity of search terms over a period of time and location. The reason why Google has normalized this data is to keep users confidentiality (the real numbers of search hits) and simplify the data extracting process. The index is shown as 0 to 100 scale, where 100 is the maximum search interest and 0 is the minimum (-to be technical- it is a ratio of a search term relative to all possible searches at the given geolocation). [↑](#footnote-ref-2)
3. http://conference.iza.org/conference\_files/WoLabConf\_2018/22651.pdf [↑](#footnote-ref-3)
4. We used OECD Migration data where it can be found at https://www.oecd.org/els/mig/keystat.htm [↑](#footnote-ref-4)
5. Although this is a good start, there are still some data limitations that constraint the possibilities of using Facebook SCI as an input to forecast migration. The biggest limitation is that Facebook data is only public for one period and that some important emigration countries like Venezuela or Syria are excluded from the sample. These constraints make it very difficult to test the true mechanisms in which this relation works. We also acknowledge that this positive correlation may be driven by other factors such as cultural, language, and geographical proximity and that a more robust approach should be taken. However, we believe that this exploratory work is a good starting point that leave us with more exciting questions and enhances our curiosity on how this type of data can, in the near future, allow effective migration forecasting. [↑](#footnote-ref-5)